MGPB

Application of the Bootstrap Method on a Large Input Data Set - Case Study on the Western Part of the Sava Depression

Rudarsko-geološko-naftni zbornik
(The Mining-Geology-Petroleum Engineering Bulletin)
UDC: 500.8: 519.2
DOI: 10.17794/rgn.2021.5.2

Original scientific paper

Josip Ivšinović[1]; Nikola Litvić[2]

[1] *Field development, INA-Industry of Oil Plc., Av. V. Holjevca 10, HR-10000 Zagreb, Croatia ORCID: 0000-0002-7451-1677*

[2] *University of Zagreb, Faculty of Mining, Geology and Petroleum Engineering, Pierottijeva 6, HR-10000 Zagreb, Croatia*

Abstract

The bootstrap method is a nonparametric statistical method that through the resampling of an input data set provides the ability to obtain a new data set that is normally distributed. Due to various factors, it is difficult to obtain many data sets for deep geological data, and in most cases, they are not normally distributed. Therefore, it is necessary to introduce a statistical tool that will enable obtaining a set with which statistical analyses can be done. The bootstrap method was applied to field "A", reservoir "L" located in the western part of the Sava Depression. It was applied to the geological variable of porosity on a set of 25 data points. The minimum number of resamplings required for a large sample to obtain a normal distribution is 1000. Interval estimation of porosity for reservoir "L" obtained by the bootstrap method is 0.1875 to 0.2144 with a 95% confidence level.

Keywords:

bootstrap; porosity; large data set; normality tests, Sava Depression

1. Introduction

Deep geological data are characterized by a relatively small set of data (<20), for which in most cases, input data for analysis is not normally distributed. The consequence of the uneven distribution of input data is a relatively small number of drilled wells in the analyzed area, lack of logging measurements, obtaining geological data from correlations with neighboring wells, etc. In the case of small oil and gas fields, very often due to complex geological structures and pronounced tectonics, hydrocarbons are obtained from smaller hydrodynamic units, which results in a smaller input data set for the analysis of geological variables. In order to obtain the most reliable data on geological variables: porosity, permeability, fluid saturation, which are crucial in the geological development of the reservoir, it is necessary to apply a reliable static tool. The bootstrap method is a method that is applicable in the case of estimating the reliability of the intervals of individual geological variables. The bootstrap method has a wide application in various branches of science (**Novoa and Mendez, 2009; Olatayo, 2013; Zhong et al., 2016; Bochniak et al., 2019; Ablanedo-Rosas et al., 2020; Phan et al., 2021; Tewari et al., 2021**). In geomathematics, and for the first time in the Croatian part of the Pannonian Basin System (CPBS), the bootstrap method was first applied by the authors **Ivšinović et al. (2021)** on the example of an oil and gas field in the Sava Depression. The authors analyzed the porosity and the cost of injection of formation water in the reservoir "K" for a small input data set.

In this paper, a set of input data for the geological variable of porosity (25 data) in the field "A", reservoir "L", which is located in the western part of the Sava Depression, is analyzed. The number of resamplings will be determined until the normality of the distribution for the input data set is obtained. The normal distribution will be tested with statistical tests of Anderson-Darling (AD) and Kolmogorov-Smirnov (K-S) after each specified number of resamplings. After determining the number of resamplings (obtaining a normal distribution of data), the interval value of the porosity of reservoir "L" will be estimated.

2. Methods

The materials and methods of this paper describe the geological setup of the investigated area, the mathematical settings of the bootstrap method, and testing the existence of a normal distribution of the data set. These analyses are needed to see the purpose and application of the bootstrap method on a large sample of data whose data are not normally distributed.

2.1. Geological settings of the investigated area

The investigated field "A" is located in the western part of the Sava Depression within the Croatian part of the Pannonian Basin System (CPBS). The area of the

Corresponding author: Josip Ivšinović
josip.ivsinovic@ina.hr

Figure 1: Position of the Sava Depression within the CPBS and the analyzed area (green). (from **Malvić et al., 2020b**)

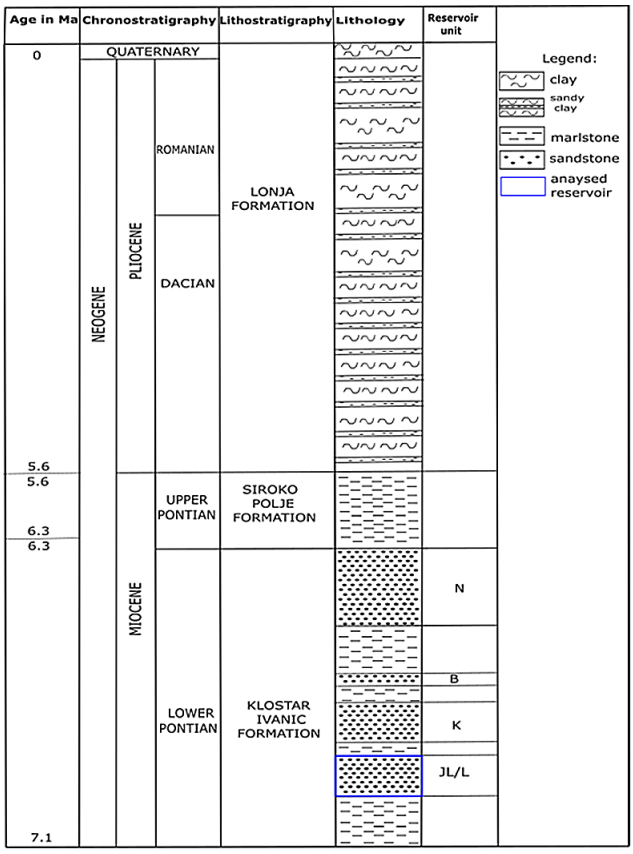

Figure 2: Geological column of field "A". (modified from **Ivšinović et al., 2020**)

western part of the Sava Depression is 8000 km^2 (**Malvić et al., 2020a**), and in the western part of the Sava Depression are the oldest and largest number of oil fields in the entire CPBS. The position of the Sava Depression and field "A" within the CPBS is shown in **Figure 1**.

Their total thickness in the deepest part of the Sava Depression reaches up to 800 meters, while it is 100-200 meters in the margins of the depression (**Vrbanac et al., 2010**). Neogene turbidites with lacustrine pelitic sedimentation formed thick heterogeneous sequences of sandstones and marls (totalling several hundreds to some thousands of metres in thickness in different depressional parts) of Upper Miocene age in northern Croatia (**Malvić, 2016**). The source of the turbidite material of today's deposits in the Sava Depression is the Eastern Alps (**Malvić, 2012**). Hydrocarbon reservoirs have been confirmed in all formations, except in the youngest Lonja Formation (see **Figure 2**). Most hydrocarbons were produced from the Upper Pannonian and Lower Pontian reservoirs in the Sava Depression.

In the research structure, hydrocarbon reservoirs were discovered in the Upper Miocene sandstones from which they are still obtained by the secondary method of hydrocarbon production. The porosity of the "L" reservoir in oil field "A" was analyzed in the paper (see **Figure 2**). Porosity data were obtained by a combination of laboratory core measurements and interpretation of logging measurements.

2.2. Mathematical settings of the bootstrap method

The bootstrap method is a nonparametric statistical method that provides an interval estimate of the value of the analyzed variable by the method of random repeated causation of the input data set. A small input set of values is considered a data set <20 (**Malvić et al., 2019a; Ivšinović and Malvić, 2020**). A sufficient number of resampling of the input set of data is an ideal statistical tool for obtaining a normal distribution of the analyzed variable (provided that the variable by its nature exhibits such a property). Ensuring a normal distribution, then it is possible to make reliable basic statistical calculations of interval estimation, expectations and variance, and parametric statistical tests. The procedure for calculating the bootstrap method is shown in **Figure 3**.

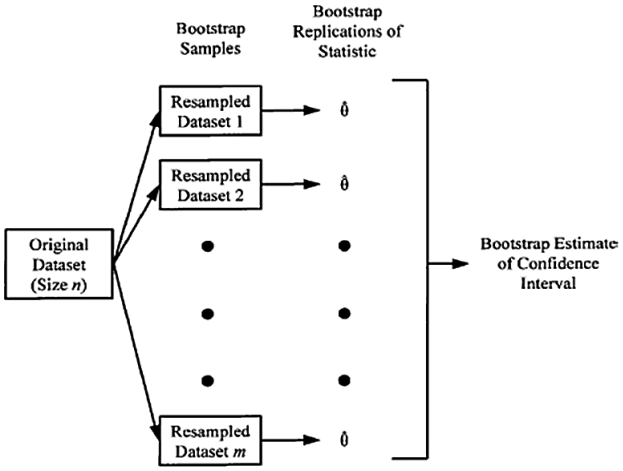

Figure 3: Bootstrap sample calculation procedures. **(Dogan, 2017)**

There are several types of bootstrap methods, and they are: Bayesian bootstrap, smooth bootstrap, parametric bootstrap, wild bootstrap, etc. In this paper, the smooth bootstrap method is applied, this method is applicable for the analysis of geological variables (**Ivšinović et al., 2021**). In the smooth bootstrap method, the input set does not change its size. Resampling randomly replaces data in a new set from the input data set. The mean value of the new data set is calculated (the same number of data remains as the original) which will be an integral part of the bootstrap data set. The number of realizations depends on the nature of the input data set.

The mean value of the resampling input data set and the bootstrap sample is calculated according to the mathematical equations described in the paper by **Ivšinović et al., 2021**. After calculating the mean value of the bootstrap sample, the standard deviation of the same sample is calculated (**Novoa and Mendez, 2009; Pajo, 2013**):

$$S_m = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\overline{X_i} - \overline{X_m}\right)^2} \tag{1}$$

Where:

S_m – standard deviation of bootstrap,
$\overline{X}_m$ – arithmetic bootstrap mean,
$\overline{X}_i$ – mean sample value after resampling,
m – number of the resampling data set.

By calculating the required bootstrap statistics (mean values and standard deviations) for the newly created bootstrap sample, an interval estimate of expectations is calculated (**Ivšinović et al., 2021**):

$$< \overline{X_m} - z\frac{S_m}{\sqrt{m}}, \overline{X_m} + z\frac{S_m}{\sqrt{m}} > \tag{2}$$

Where:

S_m – standard deviation of bootstrap,
$\overline{X}_m$ – arithmetic bootstrap mean,
z – value from the normal distribution,
m – number of the resampling data set.

The usual set reliability of the estimate of the interval is 95% (**Dogan, 2017**). The steps are repeated as many times as necessary for the input data set that is not normally distributed in the new bootstrap sample to become normally distributed.

2.3. Mathematical settings of data normality tests

In order to determine the moment of obtaining the normal distribution, it is necessary to test the data sets obtained by the bootstrap method on the normality of the data. For the control and analysis of data, the following tests were applied for the existence of a normal distribution: Anderson-Darling (A-D) test and Kolmogorov-Smirnov (K-S) test.

2.3.1. Anderson-Darling (A-D) test

The Anderson-Darling (A-D) normality test is applied when checking the distribution of different data sets. A-D test values are calculated from the following equation (**Yap and Sim, 2011; Heo et al., 2013; Jäntschi and Bolboacă, 2018**):

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\ln(\mathrm{p}_i(1-p_{n-i+1})) \tag{3}$$

Where:

AD – value of the Anderson-Darling test,
N – sample size,
p – probability.

The correction factor for the A-D normality test (AD*) for a small sample is obtained from the expression (**Yap and Sim, 2011; Jäntschi and Bolboacă, 2018**):

$$AD^* = AD \cdot \left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right) \tag{4}$$

Where:

AD^* – correction value of the Anderson-Darling test,
AD – value of the Anderson-Darling test,
p – probability.

The correction value of the A-D test for the large sample is negligible. The minimum number of test data sets is 20. The minimum "p-value" for checking the A-D test is 0.10.

2.3.2. The Kolmogorov-Smirnov (K-S)

The Kolmogorov-Smirnov (K-S) test is the most applicable statistical test for proving the normal distribution of nonparametric input data. The expression for the value of the K-S test is (**Lopes et al. 2007; Hasani and Silva 2015; Luiz and de Lima 2021**):

$$DKS = \sup\left|F(x) - P(x)\right| \tag{5}$$

Where:

DKS – value of the Kolmogorov-Smirnov test,
sup – supremum set of distances,

Table 1: Basic statistical data on porosity (parts of unit) reservoir "L"

	n	K-S	A-D	Normal distribution	Min	Max	$\bar{X}$	s
Porosity	25	Ne	0.02	No	0.145	0.239	0.202	0.026

Table 2: Normality test results for datasets after applying the bootstrap method

	m	K-S	A-D	Normal distribution
Porosity	25	No	0.02	No
	500	No	0.03	No
	1000	No	0.04	No
	1050	N/A	0.05	No
	1100	N/A	0.2	Yes
	1250	N/A	0.64	Yes

Table 3: Interval estimation porosity of reservoir "L"

	m	Confidence interval (95%)
Porosity	25	-
	500	-
	1000	-
	1050	-
	1100	<0.1875, 0.2144>
	1250	<0.1877, 0.2144>

$F(x)$ – empirical distribution function,

$P(x)$ – cumulative function of the theoretical distribution of the K-S test.

In the case of a distribution normality test, the samples are standardized and compared with the standard normal distribution. The advantages of the method are ease of application and allows the calculation of descriptive statistics for variables, which are not possible without the application of this method. The disadvantages of the method are, in the case of non-representativeness of the sample, a large expenditure of time on processing the data themselves without specific results.

3. Results and discussion

The data used in this paper are taken from a paper by **Malvić et al., 2019b**. The analyzed variable is the porosity of the reservoir "L" of the field "A". The number of analyzed porosity data set values is 25, which is a large data set. The input data set needs to be tested for distri-

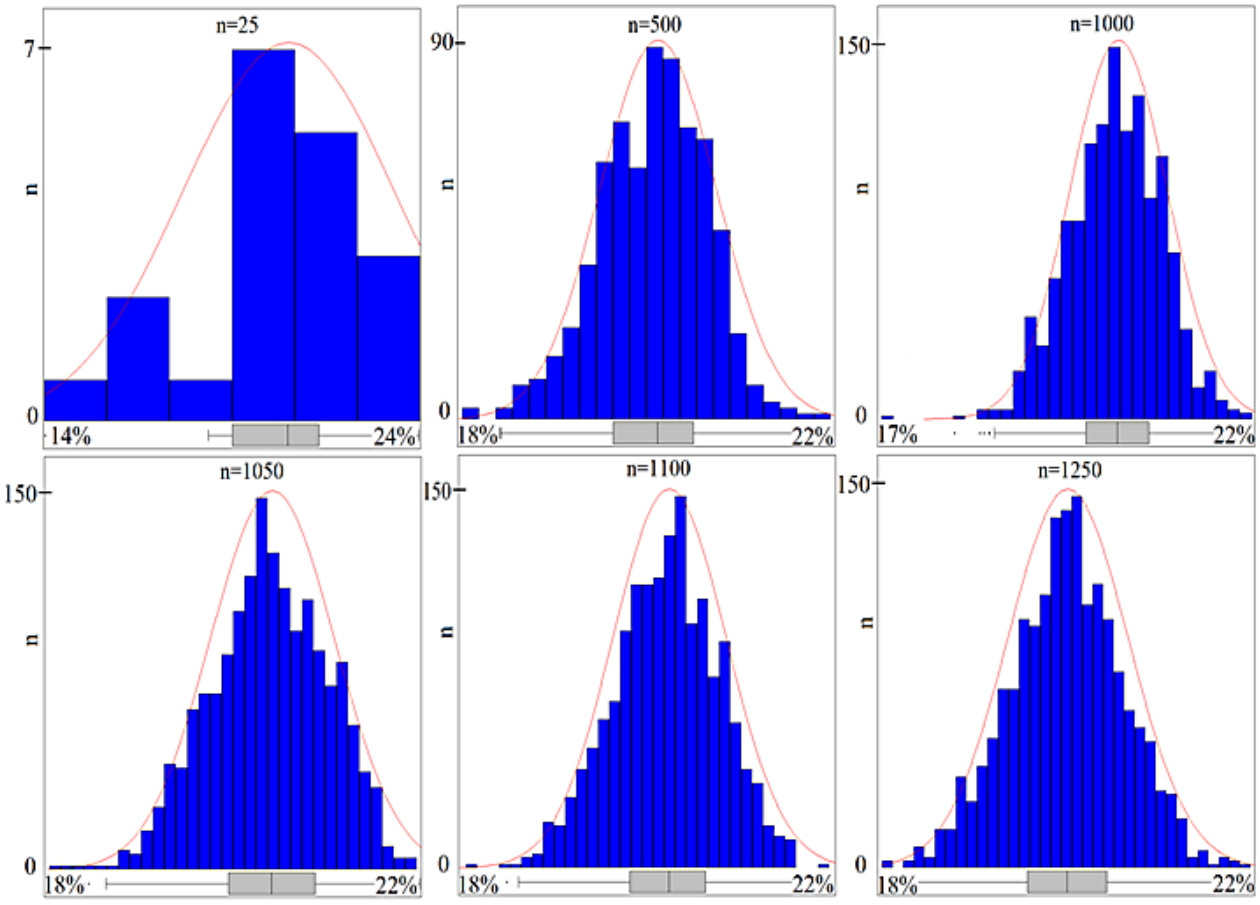

Figure 4: Histograms of porosity (reservoir "L") obtained by the bootstrap method

bution normality, in case of test negativity, it is necessary to approach the application of the bootstrap method to obtain the distribution normality. The usual number of resamplings when applying the bootstrap method is 500, 1000 and 2000 (**Carpenter and Bithell 2000; Grunkemeier and Wu 2004**). The input data on the porosity of the "L" reservoir are shown in **Table 1**.

The number of repeated resamplings applied in this paper is: 500, 1000, 1050, 1100 and 1250. The test results of the normal data distribution are shown in **Table 2**.

How can it be applied from **Table 2** since the statistical K-S test is not applicable when testing numbers greater than 1000 because no test value is obtained (test limitation of macro in Microsoft Excel)? When testing 500, 1000 and 1050, an increase in the A-D test and approaching the 0.10 limit for test acceptance is observed. After that, the value of 1250 was tested and the value of A-D increased to 0.64, which is an indication of the existence of normal distribution. An additional 1100 resamplings were tested and the test value of the A-D test was 0.20. The normality of the input data distribution for the porosity of the "L" reservoir is between 1050 and 1100 of resamplings. The calculated interval estimate of the "L" reservoir porosity expectation for resampling cases of 1100 and 1250 is shown in **Table 3**.

According to the estimate of the confidence interval of the porosity of the reservoir "L", it is visible that the difference on the fourth decimal place between the lower value of porosity for the realizations 1100 and 1250. The negligible difference between the values of the estimated intervals for 1100 and 1250 leads to the conclusion that it is not necessary to do an estimate for 2000 resamplings. A graphical representation of the results of the bootstrap method is shown in **Figure 4**.

Figure 4 shows the change in the histogram according to the appearance of the normal distribution curve (the red line). The number of classes in the case of 500 realizations is 22 (width of 0.001818 part of units), and in the case of 1000, 1050, 1100, 1250 realizations made, it is 32 (width of 0.00125 part of units). The difference between the realized realizations 1050 and 1100 is very clearly seen when there is a change in the normality of the data obtained by the bootstrap method. This can be seen from **Figure 4** how the blue columns less exceed the normal distribution boundary (red line) in cases 1100 and 1250 in which most of the blue columns are near or below the red curve. This was confirmed by the A-D test, with which the normality of data distribution is obtained after 1100 realized realizations.

4. Conclusions

The minimum amount of resamplings for a large sample on the example of the porosity of reservoir "L" is 1100. The normality of the input data was obtained between 1050 and 1100 realizations.

When testing the normal distribution of a large sample obtained by the bootstrap method, it is recommended to use the Anderson-Darling (A-D) statistical test, because the Kolmogorov-Smirnov (K-S) statistical test is not applicable to a sample larger than 1000.

Interval estimation of porosity (reservoir "L") obtained by the bootstrap method is 18.75% to 21.44% with a 95% confidence level.

The bootstrap method is applicable to a large sample, which is visible from the results of the porosity of the "L" reservoir and is therefore applicable to the entire area of the Sava Depression with similar geological characteristics as the "L" reservoir. It is used to determine the primary value of reservoir porosity and is applicable to Kloštar-Ivanić Formation reservoirs.

Acknowledgment

The authors thank the authors of Excel macro at the web address: https://www.excelforum.com/tips-and-tutorials/793135-one-sample-kolmogorov-smirnov-in-excel.html and the author of the statistical program Develve. This research has been done as part of a project funded by the University of Zagreb "Mathematical methods in geology VI" (led by T. Malvić).

5. References

Ablanedo-Rosas, J.H., Guerrero Campanur, A., Olivares-Benitez, E., Sánchez-García, J.Y. and Nuñez-Ríos, J.E. (2020): Operational Efficiency of Mexican Water Utilities: Results of a Double-Bootstrap Data Envelopment Analysis. Water, 12, 553. https://doi.org/10.3390/w12020553

Bochniak, A., Kluza, P.A., Kuna-Broniowska, I. and Koszel, M. (2019): Application of Non-Parametric Bootstrap Confidence Intervals for Evaluation of the Expected Value of the Droplet Stain Diameter Following the Spraying Process. Sustainability, 11, 7037. https://doi.org/10.3390/su11247037

Carpenter, J. and Bithell, J. (2000): Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Statistics in Medicine, 19, 1141-1164.

Dogan, C. D. (2017): Applying Bootstrap Resampling to Compute Confidence Intervals for Various Statistics with R. Eurasian Journal of Educational Research, 68, 1-17. http://dx.doi.org/10.14689/ejer.2017.68.1

Grunkemeier, G. L. and Y. Wu (2004): Bootstrap resampling methods: something for nothing? The Annals of Thoracic Surgery, 77, 1142–1144. doi:10.1016/j.athoracsur.2004.01.005

Hassani, H. and Silva, E.S. (2015): A Kolmogorov-Smirnov Based Test for Comparing the Predictive Accuracy of Two Sets of Forecasts. Econometrics, 3, 590-609. https://doi.org/10.3390/econometrics3030590

Heo, J.-H., Shin, H., Nam, W., Om, J., and Jeong, C. (2013): Approximation of modified Anderson–Darling test statistics for extreme value distributions with unknown shape

parameter. Journal of Hydrology, 499, 41-49. doi:10.1016/j.jhydrol.2013.06.008

Ivšinović, J. and Malvić, T. (2020): Application of the radial basis function interpolation method in selected reservoirs of the Croatian part of the Pannonian Basin System. Mining of mineral deposits, 14, 3, 37-42. doi:10.33271/mining14.03.037

Ivšinović, J., Malvić, T., Velić, J. and Sremac, J. (2020): Geological Probability of Success (POS), case study in the Late Miocene structures of the western part of the Sava Depression, Croatia. Arabian Journal of Geosciences, 13, 714, 1-12. doi:10.1007/s12517-020-05640-z.

Ivšinović, J., Pimenta Dinis, M., Malvić, T. and Pleše, D. (2021): Application of the bootstrap method in low-sampled Upper Miocene sandstone hydrocarbon reservoirs: a case study. Energy sources part A-recovery utilization and environmental effects, 43, doi:10.1080/15567036.2021.1883773.

Jäntschi, L., Bolboacă, S.D. (2018): Computation of Probability Associated with Anderson–Darling Statistic. Mathematics, 6, 88. https://doi.org/10.3390/math6060088

Lopes, R. H. C., Reid, I. and Hobson, P. R. (2007): The two-dimensional Kolmogorov-Smirnov test. In Proceedings of the XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research, Amsterdam, the Netherlands April 23-27, 2007.

Luiz, B. A. J. and de Lima, M. A. (2021): Application of the Kolmogorov-Smirnov test to compare greenhouse gas emissions over time. Rev. Bras. Biom., Lavras, 39, 1, 60-70. doi: 10.28951/rbb.v39i1.498

Malvić, T. (2012): Review of Miocene shallow marine and lacustrine depositional environments in Northern Croatia. Geological quarterly, 56 (3), 493-504.

Malvić, T. (2016): Regional turbidites and turbiditic environments developed during Neogene and Quaternary in Croatia. Materials and Geoenvironment, 63 (1), 39-54. doi:10.1515/rmzmag-2016-0004

Malvić T., Ivšinović J., Velić J. and Rajić R. (2019a): Interpolation of Small Datasets in the Sandstone Hydrocarbon Reservoirs, Case Study of the Sava Depression, Croatia. Geosciences, 9, 5, 201. https://doi.org/10.3390/geosciences9050201

Malvić T., Ivšinović J., Velić J., Sremac J. and Barudžija U. (2020b): Application of the Modified Shepard's Method (MSM): A Case Study with the Interpolation of Neogene Reservoir Variables in Northern Croatia. Stats, 3, 1, 68-83. https://doi.org/10.3390/stats3010007

Malvić T., Ivšinović J., Velić J., Sremac J., Barudžija U. (2020a): Increasing Efficiency of Field Water Re-Injection during Water-Flooding in Mature Hydrocarbon Reservoirs: A Case Study from the Sava Depression, Northern Croatia. Sustainability, 12, 3, 786. https://doi.org/10.3390/su12030786

Malvić, T., Ivšinović, J., Velić, J. and Rajić, R. (2019b): Kriging with a Small Number of Data Points Supported by Jack- Knifing, a Case Study in the Sava Depression (Northern Croatia). Geosciences, 9, 1, 36, 24 doi:10.3390/geosciences9010036

Novoa, C. M. and Mendez, F. (2009): Bootstrap methods for analysing time studies and input data for simulations. International Journal of Productivity and Performance Management,58,5,460-479.DOI10.1108/17410400910965724

Olatayo, T. (2013): On the Application of Bootstrap Method to Stationary Time Series Process. American Journal of Computational Mathematics, 3, 1, 61-65. doi: 10.4236/ajcm.2013.31010

Pajo, M. (2013): Bootstrap method and their application using R-programming. In Proceedings of the 1st International Conference on Research and Educatıon – Challenges Toward the Future (ICRAE2013), Shkodër, Albania, 24-25 May 2013.

Phan, T.V., Wang, G., Liu, L. and Austin, R.H. (2021): Bootstrapped Motion of an Agent on an Adaptive Resource Landscape. Symmetry1, 13, 225. https://doi.org/10.3390/sym13020225

Tewari S., Dwivedi U.D. and Biswas S. (2021): A Novel Application of Ensemble Methods with Data Resampling Techniques for Drill Bit Selection in the Oil and Gas Industry. Energies, 14(2), 432. https://doi.org/10.3390/en1402043

Vrbanac, B., Velić, J. and Malvić, T. (2010): Sedimentation of deep-water turbidites in main and marginal basins in the SW part of the Pannonian Basin. Geologica Carpathica, 61 (1), 55-69.

Yap, B. W. and Sim, C. H. (2011): Comparisons of various types of normality tests. Journal of Statistical Computation and Simulation, 81, 12, 2141-2155. DOI:10.1080/00949655.2010.52016

Zhong, H., Van Gelder, P., Wang, W., Wang, G., Liu, Y. and Niu, S. (2016): The Influence of Statistical Uncertainty in the Hydraulic Boundary Conditions on the Probabilistically Computed High Water Level Frequency Curve in the Rhine Delta. Water, 8, 147. https://doi.org/10.3390/w8040147

SAŽETAK

Primjena samonadopunjujuće metode na velikome skupu ulaznih podataka – primjena na zapadnome dijelu Savske depresije

Samonadopunjujuća metoda neparametarska je statistička metoda koja omogućuje ponovnim uzorkovanjem ulazni skup podataka za dobivanje novoga skupa podataka koji je normalno distribuiran. Zbog različitih čimbenika teško je doći do geoloških podataka u velikome skupu, a u većini slučajeva nisu normalno distribuirani. Stoga je potrebno uvesti statistički alat koji će omogućiti dobivanje skupa s kojim se mogu raditi statističke analize. Samonadopunjujuća metoda primijenjena je na polju „A", ležište „L" koje se nalazi u zapadnome dijelu Savske depresije. Primijenjena je na geološku varijablu šupljikavosti na skupu od 25 podataka. Minimalni broj ponovnoga uzorkovanja potreban za veliki uzorak da bi se dobila normalna raspodjela iznosi 1000. Intervalna procjena šupljikavosti za ležište „L" dobivena samonadopunjujućom metodom iznosi 0,1875 do 0,2144 s razinom pouzdanosti od 95 %.

Ključne riječi:
samonadopunjavanje, šupljikavost, veliki skup podataka, testovi za postojanje normalne razdiobe, Savska depresija

Author's contribution

Josip Ivšinović (PhD, research associate): had the initial idea and later on refined the idea, provided input on the design and analyses, reviewed and provided substantive feedback on the paper, and coordinated the study. **Nikola Litvić** (undergraduate student): carried out literature searches, provided input on the analyses, and edited figures and tables.